Discussion on

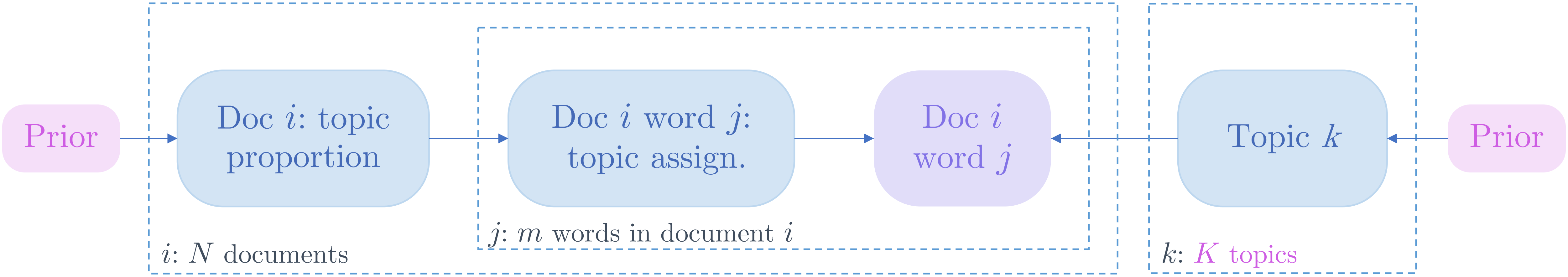
“Solving LDA Topic Models Interpretably with Near-Optimal Posterior Probability”

Adam Breuer

Hane Lee (they/them)

Department of Statistics, Columbia University

LDA and existing methods



- Bayesian mixed membership model with a generative process
- Hyperparameters, estimand, data
- Maximize the "posterior"
  - Variational inference: EM algorithm, Gibbs sampling

Challenges of conventional LDA methods

- Choosing the number of topics K
  - Too small: miss themes
  - Too big: topics overlap, become too specific
- Sensitivity to other hyperparameters
  - Priors, number of iterations
- Interpretability of topics
  - Often involves manual labeling, evaluation
  - Sometimes topics don't make sense
- Instability
  - Models generated with the same parameters over the same data will produce different results

Proposed method (Breuer)

- Precompute a large set of candidate topics (~20,000) using co-occurrence
  - Given each topic label, compile a set of words that co-occur in documents
- Fit an LDA model incrementally using a greedy algorithm
  - At each iteration, the topic that maximizes the posterior probability is selected and added to the solution set
  - With every additional topic, increment to posterior probability is positive but decreases.
  - Combinatorial set selection problem, not a gradient descent problem
- Algorithm "provably obtains the near-optimal topic from which each word in the dataset was probably drawn."

Contributions

- Improvement in topic interpretability
  - Topics are pre-generated from labels using co-occurrence
  - Pre-generated topics can be tested for standard interpretability criteria
  - Topics are less likely to overlap
  - Intuitive explanation of topic selection process
- User does not choose topic sparsity priors, use uniform priors
- Algorithm is deterministic
  - Given same input, same output
- Logarithmic computation time
- Framework for causal inference

Discussion

- How (non)trivial is topic generation?
- Hyperparameter κ
  - Cardinality constraint: (# of topics: K) $\leq \kappa$(# of Documents)
  - Average number of topics linked to each document is bounded by κ
  - Smaller κ → fewer topics, sparser solution
  - Larger κ → larger posterior probability
  - Sensitivity analysis, decision criteria
- Topic concentration on certain documents/words?
- Evaluation criteria other than coherence

Discussion on

“Boundless but Bundled: Modelling Quasi-infinite Dimensions in Ideological Space”

Philip Warncke, Flavio Azevedo

Hane Lee (they/them)

Department of Statistics, Columbia University

Context

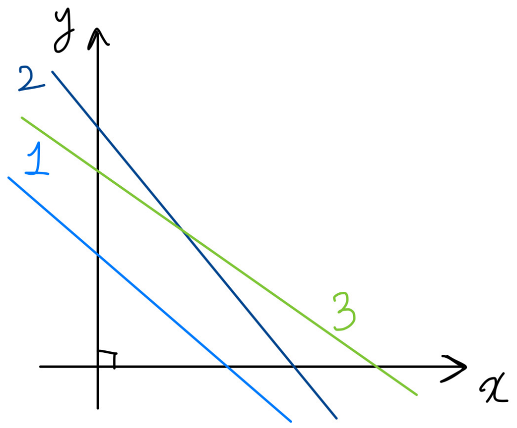
- Review on ideological dimensionality
  - Researchers have often assumed a unidimensional spatial model (left-right) when operationalizing ideology
  - Unidimensional model implicitly up-weights responses that conform to the liberal-conservative division
  - Multidimensional models are also common
  - Multidimensional model down-weights responses that conform to the liberal-conservative division
- Among analyses in literature, number of selected issue items correlates strongly and positively with the number of ideological dimensions used

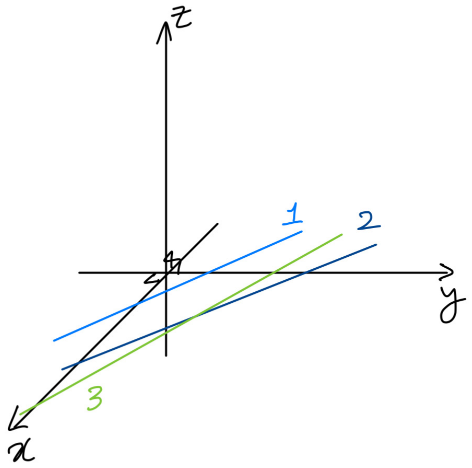
Optimal Number of Latent Ideological Dimensions

- Exploratory Graph Analysis (Golino et al. 2017)
  - Finds optimally sparse representation of the item correlation matrix using LASSO
  - Weighted community detection on the matrix while minimizing total information entropy
  - Number of communities detected: optimal number of dimensions
- Using EGA, authors show that optimal number of dimensions goes to infinity as the number of issues increases

Optimal Number of Latent Ideological Dimensions

- Q: Are these dimensions orthogonal?
  - Given highly correlated, overlapping items, EGA may still produce separate factors (Golino and Epskamp 2017)
  - Each factor may not necessitate another dimension for explanation
  - Are these correlated "dimensions" necessary?
- Q: How much explanation does each dimension add?
  - Ex. NOMINATE with 2 orthogonal factors. First dimension explains 83%, the second only 3%.





Bayesian Hierarchical Model of Ideology

- "Virtually all latent dimensions identified in policy position data are strongly and consistently positively correlated with one another."
- "Although complex enough to warrant separate spatial representation, all latent ideological dimensions seem to be tethered to an overarching, yet somewhat imprecise, uni-dimensional origin."

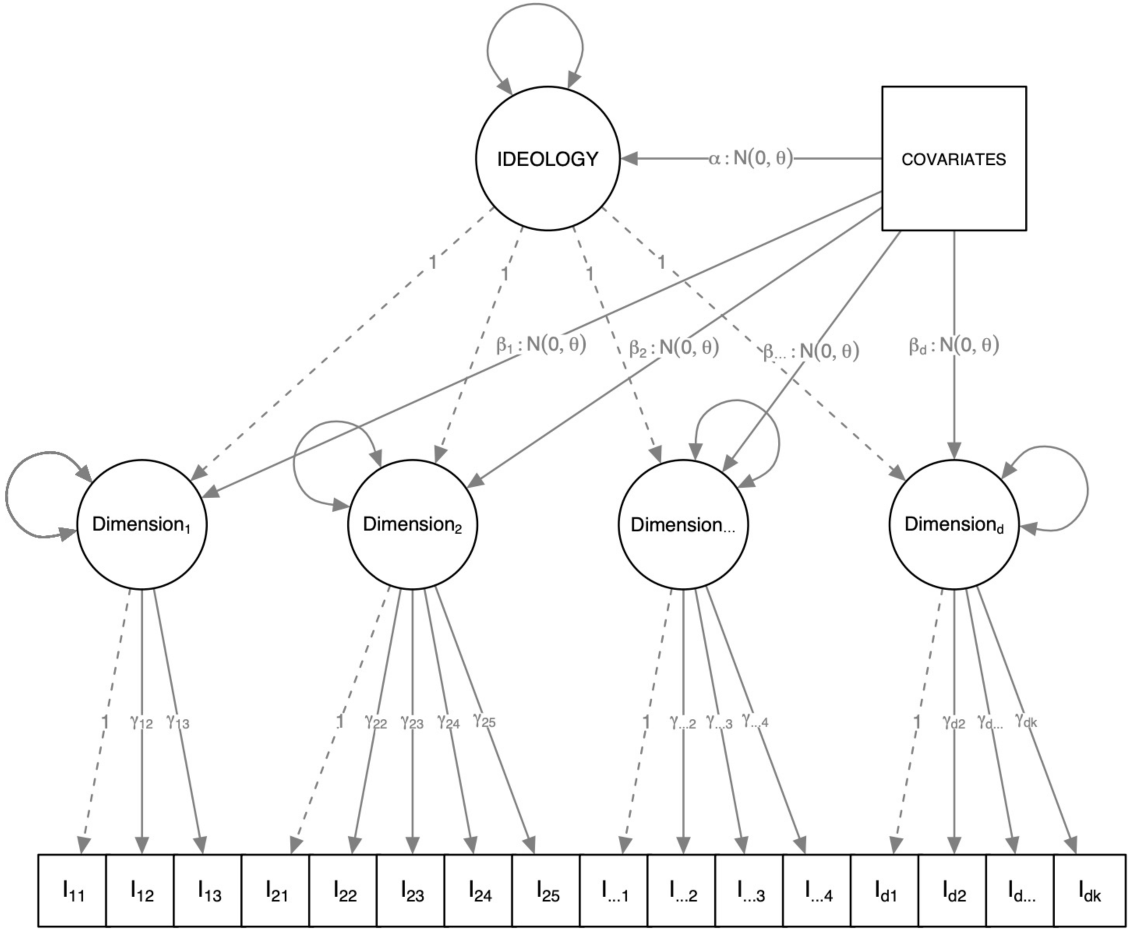


Figure 5 of paper

Bayesian Hierarchical Model of Ideology

- 6 identified sub-dimensions from ANES items: "1) poverty reduction, 2) New Deal issues, 3) socio-cultural issues, 4) racial justice, 5) moral & sexual chauvinism, and 6) anti-immigrant chauvinism."
- Which covariates predict which ideological sub-dimension?
- Q: Causal (mediation) interpretation?
  - Mediation assumptions: 1) $Y_i(x', m), M_i(x) \perp X_i = x$ 2) $Y_i(x', m) \perp M_i(x) | X_i = x,$

Policy issues and ideology

- Q: Are policy issue positions ideological?
  - Ex. Racial "ideology" and sexual chauvinism
- Q: Given that ideology is involved, to what extent are policy issue positions explained by ideology?
  - Ideology may not be the sole determinant of issue positions
  - Many issue positions are correlated with partisanship in the US
  - Ex. There have been efforts to distinguish symbolic racism from conservative ideology (Zigerell 2015)